**Preparation Info for Hackathon 2017**

**November 18, 2017**

Bring laptops to work on. You may use any software.

**BUSINESS CASE**

A one-page business case will be provided at the beginning of the competition.

**DATA**

The data and a data dictionary will be provided at the beginning of the competition. The data consists of 8000 rows and 21 variables, one of which will be a binary target variable. Use the remaining variables build a predictive model of the target. The data is real and not terribly dirty, but will benefit from cleaning,

A single data file will contain the data with the target value for building models, and the data without the true target, for testing the model. A variable named **Sample** will be included to easily distinguish each case. For the test set, the Sample variable value will be “Test”. This will allow you to perform any data cleaning and manipulation on both data sets simultaneously with needing to have separate files for each, so that the models you build will be applicable to the test data. To reiterate, *any records that are in this test sample have the target variable missing*.

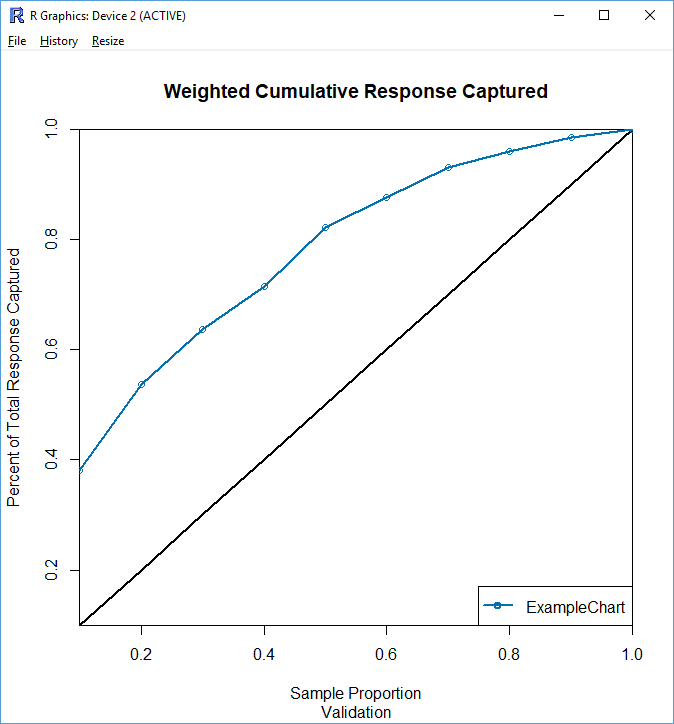
The Sample variable for the data rows with the known target for model building will be randomly assigned “Estimation” or “Validation” with a 50/50 split. This is done to facilitate your own modeling on a subset of the data set (the Estimation set) and testing how well it does with a known target on another subset (the Validation set). However, you may ignore this distinction and use the data with the known target however you like. The data will be arranged as below, but with more rows and predictor variables (and different names for the target and predictors) of course.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Target** | **Pred1** | **Pred2** | **Sample** |
|  | Yes | 1 | a | Estimation |
|  | Yes | 2 | a | Validation |
|  | No | 3 | b | Validation |
|  | Yes | 3 | c | Estimation |
|  | No | 5 | d | Estimation |
|  | No | 1 | a | Validation |
|  | Yes | 2 | b | Validation |
|  | No | 3 | b | Validation |
|  |  | 6 | c | Test |
|  |  | 5 | c | Test |
|  |  | 2 | a | Test |
|  |  | 1 | d | Test |

You will add a new variable to the data set labled “Score” which have the predicted probability, from your best model, of a positive target value for each row.

**MODEL SCORING**

Models will be assessed by assigning probability scores of a positive target, generated from your best model, to data where the true target variable is unknown to competitors. This scored data will be submitted to judges at the end of the competition, who have access to the known true target. The probability scores and the true target will be used to generate a cumulative captured lift chart. The final value for comparison will be the percent captured at the 40% sample proportion level.



An example of a “cumulative captured” lift chart, and the 40% captured level. This model is capturing about 72% of the positive responses.

Competitors do not need to generate lift charts. You may use any model quality measure you are comfortable with, but all models will be assessed with this method.

Save the scored dataset as a “.csv” file, with the filename “*your teamname.csv*”. e.g, if you are using R,

write.csv(Dataset, “C:/yourdirectories/…/yourteamname.csv”)

Keep in mind that you will have to present, explain and defend your model and your approach to the business problem!

**TROUBLESHOOTING**

Two or three instructors will be available to troubleshoot any technical problems, but will be reluctant to give advice beyond that.

**FINAL SUBMISSION:**

Final scored dataset is due at 2:15 p.m.

**MANAGERIAL PRESENTATION:**

Starting after submission you will prepare a five-minute PowerPoint presentation that answers the questions “so what?” and “what next?” This will describe how you will use the results and what further work needs to be done (if any) in order to use the results, WITHOUT TECHNICAL JARGON. The presentation will be followed by, at most, 10 minutes of Q&A. Teams will be judged on communication (clarity, logic, persuasiveness), business understanding (tactical and strategic considerations), and, since *questions* are not restricted by the NO TECHNICAL JARGON rule, on technical competence. Presentations will begin at 3:15. Sequence will be assigned with more junior teams first. Presentations will only be to the judges, and to teams who have already presented. (Thus the first team will present only to judges, and the last team will present to everyone).